#Open Refine

Open Refine, formerly known as Google-Refine, is an open source data cleaning & data transformation tool that provides a user-friendly interface for cleaning, transforming & enhancing messy or unstructured data.

Useful to identify similar redundant rows & merge them into clusters.

#Image labelling. PIL

Python Imaging Library, is a popular library for basic image processing tasks, but it has been succeeded by the Pillow library, which is a more actively maintained and enhanced fork of PIL.

Image opening & saving.

Image manipulation

Basic Drawing.

Color & channel manipulation

Week 4

MODEL THE DATA.
Which software to use when & for which purpose!

Date :
Page No.

- **Excel** . for i) Correlation.
  (How often is there an effect
  Eg: How often increase in Math marks imply increase in Phy marks?)

  ii) Regression.
  (How much is there an effect?)

  iii) Outlier Detection.
  (When does this relation fail?)

- **Python** for. i) Classification
  (Given data pt, which grp should it belong based on the history data?)

  ii) Forecasting.
  (Future prediction)

  iii) Clustering
  (can we group things into similar entities?)

- Others
  - ▶ R/Rstudio — one of the most powerful tools for data modeling
  - ▶ Rattle — non-programmatic R-based application that helps you analyze data
  - ▶ Py caret. similar to Python; help you build models and explore models fairly, quickly.

* Correlation (using Excel). Use Data Tab > Data Analysis > Correlation. (Add if not there using Add-ins)

- Interpret correlation matrix.

	A	B	C
A	1		
B	ab	1	
C	ac	bc	1

$1 > (ab) > 0$ +ve correlation

slope of linear map: $ab \to 1$ high/strong +ve corr; $ab \to 0$ weak +ve corr.

$-1 < (ab) < 0$ -ve correlation; $(ab) \to -1$ strong -ve corr; $(ab) \to 0$ weak -ve corr.

- Use scatterplots to visualize each correlation

* Regression (using Excel).

Use Data Tab > Data Analysis > Regression. (Add if not present)

single linear / multiple linear (automatic selection based on selected data containing 1 or >1 indep var).

- Single linear Regression.
  1 indep var.
  1 dep var.
- Multiple linear Reg.
  (>1) indep; 1 dep.

SUMMARY OUTPUT

:

Mult R Square 0.884585151 → for single linear regression

R square 0.7824...

Adj R sq 0.7740... → for multiple linear regression

	Coeff.
Intercept	c_1
x_1 City Pop.	d_1
x_2	d_2
:	:

$y = c_1 + d_1 x_1 + d_2 x_2 + \ldots + d_n x_n$.

* Outlier Detection.

	A	B
1	~~Variable~~ Values	Outliers
2	53040	= FALSE = OR(A2 < D4, A2 > D5)
3	23671	FALSE
4	255215	TRUE.

Date :
Page No.

	C	D
1	Q1	3453 = QUARTILE.EXC(A:A, 1)
2	Q3	100158 = QUARTILE.EXC(A:A, 3)
3	IQR	96705 = D2 - D1
4	LowerB	-(...) → 0. (can't be -ve)
5	UpperB	245215.5

= D1 - (1.5 * D3) → = D2 + (1.5 * D3)

(Q1 - 1.5 IQR , Q3 + 1.5 IQR).

Model the Data : Pycaret library, Shap library.

Pycaret is an open source ly library designed to simplify the end-to-end machine learning workflow. It was created to help data scientists & machine learning practitioners streamline the process of building and deploying machine learning models by automating many of the common tasks & reducing the need of extensive coding.

```
import pycaret
from pycaret.datasets import get_data
index = get_data('index')
  # displays the 30 different datasets present in
                    # pycaret_dataset
```

juice.shape
returns the dimensions of the dataset

random seed. to ensure some randomization

Date :
Page No.

```
from pycaret.classification import *
clfi = setup(juice, target = 'Purchase', session_id = 123,
             log_experiment = True, experiment_name = 'juice1',
             normalize = True, feature_selection = True)
```

we setup an experiment for pycaret to use
this data and get a sense on what kind
of model are we dealing with.

The output you get; based on this you can have
quick analysis of for further process
Plug & play Pycaret.

```
training_data = get_config('X_train')
# we receive the training dataset
training_data.shape
models()
# shows the details of models performed on the
# dataset, and how well is it to be chosen.

best_model = compare_models()
best_model
# shows the best-fit model.
# Either use this or the user may explicitly
# use some other model of their choice
# say lr = create_model('lr')
```

tuned_best_model = tune_model(best_model)
# Fine tunes the hyperparameters of the model to
# establish better performance

Date :
Page No.

plot_model(tuned_best_model)
# displays ROC curve
# Receiver Operating Characteristic
plot_model(tuned_best_model, plot='confusion matrix')
plot_model(tuned_best_model, plot='feature')
# different plot curves.
interpret_model(tuned_best_model)
# shows the graph with shaft values.
save_model(tuned_best_model, model_name='best_model')
# saves the model as a pickle file
# extension = .pkl.

* Pycaret helps a lot in automating Machine learning.

Clustering with Python.

import pandas as pd
import numpy as np
#clustering method KMeans gets imported
from sklearn.cluster import KMeans

stockData = pd.read_csv('https://raw.githubusercontent.com/~/...csv')

StockData.head()

```
# Multiplying those % columns that range from 0 to 1
stockData['Free Float ...'] = stockData['Free...'] * 100
stockData['RoE %.'] = stockData['RoE %.'] * 100
  :
stockData.describe()

features = stockData.columns[2:]
# storing features.
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
stockDataFeatures_scaled = scaler.fit_transform(stockData[features])
stockDataFeatures_sorted = pd.DataFrame(stockDataFeatures_
                                        scaled,
# Scaling the DB                        columns=features)
stockDataFeatures_scaled.describe()



stockDataFeatures_scaled

kmeans = KMeans(7) # stores the model object.
clus = kmeans.fit_predict(stockDataFeatures_scaled)

stockData['cluster'] = clus
stockData['cluster'].value_counts()

clusterDesc = pd.DataFrame(stockData.iloc[:,2].groupby('cluster')
                                                .mean
                                                .round(3))
clusterDesc.insert(0,'size',stockData['cluster'].value_counts())
```

• What is a shapefile and how to read a shapefile using python?

A shapefile (→ extension (.shp)) is a widely used geospatial vector data format for Geographic Information Software (GIS).

A shapefile stores geometric & attribute data for geographic features, such as points, lines & polygons & is often used to represent geographic info like maps, boundaries & locations.

```
import geopandas as gpd
shp_path = 'path/to/your_shapefile.shp'
gdf = gpd.read_file(shp_path)
# printing head...
gdf.head()
```

Geospatial Analysis also known as spatial analysis or geographic information system analysis (GIS), is the process of examining & interpreting data that has a geographic or spatial component.

W5
Deep learning Models

Date
Page

#code.
For this example, we are checking ~~there~~ whether
there is a need of a new store near The Empire
State Building
[Geospatial analysis :
The objective of using geospatial analysis is to make decisions about opening up new restaurants (or retail stores, bank branches, airports etc). Here in this example, we will use a dataset consisting of Starbucks & McDonald's locations in New York.]

```
import pandas as pd
import numpy as np
import folium
import geopy.distance
#Read file containing location details
df = pd.read_csv('https://drive.google.com/....')
df.head(10)
```

	lat	long	store	address
0	40.7	-74.008	Starbucks	375 Pearl st, New York, NY 10038
1	⋮	⋮	⋮	⋮

```
# Merge lat & long columns to help us
df['coordinate'] = '(' + df['lat'].astype(str)
                   + ',' + df['lng'].astype(str) + ')'
```

Date:
Page No.

```
df.head(10).
```

```
# Search from web : The lat-long of the target
# location
# Empire State Building : (40.7488, -73.9852)
NY_coord = (40.748488, -73.985238)

# Step 1: Compute distances from the stores
~~distance = []~~
distances_km = []
for row in df.itertuples(index = False):
    distances_km.append(geopy.distance.distance
                        (NY_coord,
                         row.coordinate).km)
    df['Distance'] = distances_km
df.head(10)

# Visualize Data on map (Step 2)
# Empire State building coordinates
m = folium.Map(location = [40.748488, -74.985238],
               zoom_start = 10)
```

(geopy = library, distance = method)

Place markers for stores in the map. → returns (ind, row) enlarge

```
for i, row in df.iterrows():
    lat = df.at[i, 'lat']
    lon = df.at[i, 'lng']
    store = df.at[i, 'store']
    if store == 'McDonalds':
        color = 'red'
    else:
        color = 'green'
    folium.Marker(location=[lat, lng], popup=store,
                  icon=folium.Icon(color=color)).
                  add_to(m)

m. # displays the map
```

Step 3 # Compute no of stores located in a given radius

```
# All stores in a radius of x = 10 kms
df.loc[df['Distance'] > 10]
```

'loc' accessor that allows to select data from a DataFrame by label-based indexing.

Step 4 : Compute the closest & farthest store from # target_loc.

df_farthest = df.iloc[df.groupby('store')['Distance'].idxmax()]

df_closest = df.iloc[df.groupby('store')['Distance'].idxmin()]

Date : .agg(pd.Series)

Page No.

df_new = pd.concat([df_farthest, df_closest])

df_new # prints the dataframe with closest & farthest

Visualize the farthest & closest using folium
similar to what has been before

Sentiment Analysis. with Excel & Azure ML

* Azure ML. Azure ML

Open Excel.

Insert > My-add-ins > Azure ML > Text sentiment analysis

Then select data

Select output data

* TextBlob library

```
import pandas as pd
import numpy as np
import spacy
from textblob import TextBlob
from sklearn.metrics import classification_report
# importing all neccessary libraries.
```

```
data = pd.read_csv('....csv')
data.head()  # reading csv files & showing data.head()

data['TextBlob_Subjectivity'] = data['review'].apply(lambda x:
# applying TextBlob_Subjectivity formula          TextBlob(x).sentiment.
                                                  subjectivity)

data['TextBlob_Polarity'] = data['review'].apply
# applying TextBlob_Polarity formula (lambda x: TextBlob(x).sentiment.polarity)

data['TextBlob_Analysis'] = data['TextBlob_Polarity'].apply
# applying TextBlob_Analysis formula.  (lambda x: 'negative' if x<0
                                        else 'positive')

data

print(classification_report(data['sentiment'],
                            data['TextBlob_Analysis']))
```

Image classification.
with Keras, with Google Cloud Platform,

- Keras: Keras is an open-source deep learning framework written in Python. It is designed to be user-friendly, modular and extensible, and making it accessible for beginners while providing advanced functionality for experienced practitioners.

```
from keras.models import Sequential
from keras.layers import Dense

# Define sequential model
model = Sequential()
# add layers to the model
model.add(Dense(units = 64, activation = 'relu', input_dim = 100))
model.add(Dense(units = 10, activation = 'softmax'))

#compile model
model.compile(loss = 'categorical-crossentropy', optimizer
                = 'sgd', metrics = ['accuracy'])

# Train the model
model.fit(x_train, y_train, epochs = 10, batch_size = 32)
```

Date :
Page No.

W6
Design your output

Visualizing forecasts with Excel.

Given a data series (time-series) of currency (many) value fluctuations

- Trendlines ? 'Insert' → 'Line Chart'.
  (Sparkline

- correlation visualization := Scatter Plots.
  (Add trendline to understand the linear ∞ regression relation)

Date :
Page No.

- Correlation Matrix — 'Data Analysis' → 'correlation'.

- colour grading based on values — 'Conditional Formatting' → 'Color Scales'

Visualization of animated data with Flowrish.

Use 'Bar Chart Race'

Kumu.

(Primarily) to visualize ~~[illegible]~~ Inter linkages.

105

Design your output.

Design the output

General

1. Excel
2. Google Data Studio
3. Power BI
4. Tableau

Specialized

1. Excel (VBA)
2. Flourish Studio
3. Kumu
4. QGIS.

Flourish Studio → animated visualizations

Kumu → useful for network diagram

QGIS → Geospatial analysis

Visualization with Excel

- ✓ Pivot Table
- ✓ Trends using Sparklines
- ✓ Correlation matrix.
- ✓ Charts from pivot table.

Visualization with Powerpoint.

- ✓ Use pivot table and then scale the unit such that it can taken as the width of a box in ppt.
- ✓ Insert box & textbox.
- ✓ Have a slide for each year/category
- ✓ Use transition 'Morph' & do some editing.

Visualization with Kumu

Kumu is a web-based platform designed for visualizing & exploring complex systems, & relationships.

Date :
Page No.

Key features : Relationship Mapping
Data Import & Integration
Visual customization
Dynamic Mapping.

Visualization with Flourish Studio.

Flourish Studio refers to a web-based data visualization platform called Flourish. Flourish is designed to make it easy for users to create interactive and engaging data visualizations without requiring extensive coding or design skills.

- ✓ User friendly interface
- ✓ Templates & designs
- ✓ Data import & connectivity
- ✓ Interactivity.

(W7) Narrate a story

Numbers — suppose usage of some numerical figures and statistics, percentages, etc

Visuals — Charts, trendlines, etc
Animated charts.

Text — Descriptive statistics

Illustrations — Pictographical representation. (canvagen)

- narrativeschool.com/quill/tableau/free-trial

(W8) Deploy the results

Anonymize the data	Build app	Host app
• Tools (ARX, Amnesia)	• Notebook (colab, kaggle)	• Content (Github, Desktop)
• Libraries (Faker, nimesis)	• Data app (streamlit, shiny)	• Apps (Heroku, Glitch)
	• Web app (Flask, Tornado)	• Infra (AWS, Azure)

Secure App ⟶ Scale App

Libraries to build web applications

- **Streamlit** Streamlit is an open source Python library that is used for creating web applications for data science & machine learning. It is designed to be easy to use, and allows developers to create interactive and customizable web apps with minimal effort.

Basic Streamlit components

1) Text
2) Interactive widgets
3) Charts & plots
4) Charts & Plots (matplotlib.pyplot)
5) Dataframe display (pandas)
6) Interactive maps (folium lib)

ngrok

ngrok is a tool that creates secure tunnels to localhost, allowing you to expose a local server to the internet.

Services to host web application.

Heroku Heroku is a cloud platform as a service (PaaS) that allows developers to deploy, manage and scale applications easily. It provides a platform for building, deploying, and scaling web applications without the complexity of managing the underlying infrastructure.

22nd Dec
- TDS pyq
- MLP pyq
- MLP session

23rd Dec
- MLP, TDS

Date : revision
Page No.

22/12/23

PYQ TDS Apr 2023

Q1) The dataset consists of geographic, demographic information about countries & their respective GDPs. You would like to visualize this data & study the relationship between the location of countries & their GDPs. You decide to use PowerBI to visualize the dataset. But you would also like to generate summary of the data.

Suggest an appropriate way.

1) Map visualization
2) Scatter plots
3) Treemaps
4) PowerBI Q&A
5) Card Visuals
6) Key Influencers

Q2) Your project requires you to study the districts & their respective health indicators. You have a shapefile with you that provides the required details. The objective of the project is to identify and carve out districts that present high levels of health indicators. How could QGIS be used here?

QGIS : Quantum Geographic Information System.
It is a software, open source software, that allows users to view, edit & analyze geospatial data.

— Here QGIS can be used to create shapefiles for districts with high levels of health indicators.

Q3) What are the two outputs provided by Excel Azure Machine learning plugin?

Sentiment, Score.

Q4) The dataset consists of year, annual cotton production, annual rainfall, loan interest rates & fuel prices. You would like to compute the correlation coefficient between annual cotton production & other variables in the dataset to analyse the effects of various variables on the target var.

Using Excel, it is a wonderful way to get correlation coefficient.

Q5) You would like to prepare your dataset before analysis. You choose pandas-profiling library to perform explanatory analysis.

Choice is appropriate

Pandas profiling provides information about outliers

96) Comicgen is a useful tool in narrating Data stories using comics. Which of the following is a function of comicgen?

Date :
Page No.

✓ comicgen creates comic characters
✓ Comicgen provides options to custom create different comic characters & their emotions and pose.
✓ Comicgen can be easily integrated into Google sheets or Excel to narrate your data stories.

97) A very large Matrix A has a lot of zeros in it. which function from scipy library is useful for efficient storage of A?

csr-matrix → compressed sparse row matrix.

98) Which library has functions & tools that are useful in analysis of large graphs?

scikit-network.

99) Kumu is a tool that allows you to:

visualize complex network data.

Q10) Which library is used to extract data from wikipedia pages? wikipedia

Date :
Page No.

Q11) A dataset is provided to you about countries & respective populations. You plan to visualize the data in Tableau using map representation But you are unable to do so because the map representation is not activated for you to choose. What could be the issue?

col-Name	col-type
Country	String
Population	Integer

There might be column type incompatibility issues

Q12) _____ is helpful to understand the structure of a website before writing a scraping script.

Developer Tools

Q13) requests library has tools to get a webpage's html contents into Python.

Q14) type of location can be retrieved using Nominatim in Python. TRUE

Q15) Is there any restriction to the type of delimiter that can be used in text-to-column func^n in Excel? No, no restriction

Q16) What is the y-axis in autocorrelation plot?

correlation

x-axis in autocorrelation plot is time lag.

Q17) Which tools cannot be used for anonymising the data? PowerBI.

Q18) For a one-time anonymization, static anonymization is sufficient. TRUE

Q19) Variable X has values cc, AA, BB, AA. This info is represented as shown below:

AA	BB	cc
0	0	1
1	0	0
0	1	0
1	0	0

pandas.get_dummies

Q20) K-means is typically influenced by the start values. What option in sklearn.cluster.KMeans help reduce the impact?

n_init.

Q21) Pandas dataframe has salary-range
high
medium
medium
low
high
You are interested in finding out how many employees are present in each category?

Date :
Page No.

DF['salary-range'].value-counts()

Q22) We have an imbalanced dataset in class. Which feature in DecisionTreeClassifier() will help us tackle the problem?

class-weight.

Q23) $\sum_{i=1}^{100} \frac{|y_i - \hat{y}_i|}{100} = ?$ sklearn.metrics.mean-absolute-error

Q24) We are interested in fitting an ARIMA model to our time-series data. Specifically we are interested in a moving average model of 0, setting a lag value of 4 for autoregression, & a difference order of 1.

ARIMA(..., order(4, 1, 0))

- ARIMA → AutoRegressive Integrated moving Average.
- 4 → lag value
- 1 → diff order
- 0 → moving avg model

Q25) pycaret is a ?

pycaret is a low-code machine learning library.

Date :
Page No.

Q26) 'subjectivity' & 'polarity' are two properties returned by sentiment function of which library?

TextBlob.

Q27) What is the purpose of 'subjectivity' & 'polarity' ?

'subjectivity' score ranges from 0.0 to 1.0

- 0.0 → very objective
- 1.0 → very subjective

'polarity' score ranges from -1.0 to 1.0

- -1.0 → negative sentiment
- 0.0
- 1.0 → positive sentiment

Q29) What does classification-report function from the sklearn.metrics module?

It builds a text report displaying the main classification metrics

Page No.

Q30) csr-matrix from scipy library

helps reduce matrix space when there are a lot of zero entries in the matrix

Q31) Google Studio is a tool that allows you to create dashboards for small scale projects

Q32) what is the name of the tab that is used to identify API calls in the Inspect element in any browser?

Network

Q33) Which of the following libraries are used to construct API urls? urllib

Q34) A time series with a significant autocorrelation at lag 1 implies that

current observation is correlated with previous observation.

Q35) Sentiment analysis function in TextBlob

calculates a numerical score for the sentiment of the text.

Q36) Match the following

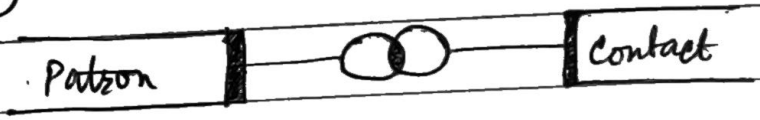
1. Worksheet — a) contains a seq^n of worksheets or dashboards that work together to convey information

2. dashboard — b) contains single view along with shelves, legend, data pane.

3. story — c) collection of view from multiple worksheets

1 → b 2 → c 3 → a

Q37) Identify the join used below:



Inner join.

Q38) Which attribute does not belong to the category of data while categorizing data based on specifics of the disclosure risks from which a dataset is to be protected?

non-identifying attributes.

Q39) What is one of the ways/methods of removing outliers?

IQR analysis

Q40) On applying OHC, what happens to the dataframe the # columns increases.

Q41) _____ table in Excel is a way of quickly summarizing your data.

Pivot table

Q42) Fill the missing line.

```
distances_km = []
for row in df.itertuples(index = False):
    distances_km.append(
        ... <missing line>
    )
df['distance'] = distances_km
df.head(10)
```

```
geopy.distance.distance(location1, location2).km
```

Q43) Write a piece of code to display details of 9 scheduled airlines using BeautifulSoup.

```
import requests
import pandas as pd
from bs4 import BeautifulSoup
website_url = requests.get('.....').text
soup = BeautifulSoup(website_url, 'html.parser')
required_table = soup.find_all('table')[0]
```

```
df = pd.read_html(str(required_table))
df = pd.DataFrame(df[0])
df
```

Date :
Page No.

Q44) Write the correct format of logical expression in Tableau.

```
IF [Sales] > 1000 THEN 'High Sales'
ELSE 'Low Sales' END
```

Q45) What is the attribute to get latitude & longitude of a location.

```
Nominatim().geocode('location_name').latitude
                                    .longitude
```

Q46) In streamlit, how to create a dropdown menu select for Hobbies having 'Dancing', 'Reading' 'Sports' as options.

```
import streamlit as st
st.selectbox("Hobbies:", ['Dancing', 'Reading', 'Sports'])
```

Q47) What is the function of Scikit-network package?

- social network analysis
- analysis of large graphs

Q49) Write the name of the library from which the following methods originate:

i) find() — BeautifulSoup
ii) get() — requests
iii) find_all() — BeautifulSoup
iv) prettify() — BeautifulSoup

Q50) We are analyzing how much the no of lecture hours attended by students affect their exam scores. Which excel function would you use as a starting point in this analysis?

SLOPE()

Q51) How to carry out a regression analysis in Excel?

Data Analysis Toolpak.

- What is the function of scikit-network.
  - social network analysis
  - analysis of large graphs.
- How to get distance in km b/w two locations using geopy module?

```
geopy.distance.distance(loc1, loc2).km
```

- OO Term? Inner Join.